RECURSIVE PREDICTIVE ABSTRACTION FOR UNSUPERVISED TEMPORAL PARTONOMY INDUCTION

Anonymous authors
Paper under double-blind review

ABSTRACT

We present **RePart**, an unsupervised framework for inducing temporal partonomies through recursive predictive abstraction. RePart couples a shared causal predictor with Bayesian composition evidence over current-level representations and prediction residuals, recursively merging adjacent units into increasingly coarse events without predefined hierarchy levels. The hierarchy is induced through discrete Bayesian composition, while predictive and composition-aware objectives shape the representation. On Breakfast Actions, 50Salads, and Assembly101, RePart achieves the best unsupervised hierarchical F1 on all three benchmarks and improves over PARSE on every TED-Sim/hF1 comparison, while remaining competitive on fine- and coarse-grained segmentation. Ablations show that recursive Bayesian composition provides a strong structural inductive bias, with joint state–residual evidence and representation adaptation further refining higher-level organization. In controlled Assembly101 action anticipation experiments, partonomy-conditioned models outperform fixed-scale models, with RePart remaining strong across increasing prediction horizons. These results show that recursively induced predictive structure provides coherent part–whole organization and useful context for downstream temporal reasoning. Code is in the supplementary and will be released publicly after review.

1 INTRODUCTION

Understanding temporally extended activities requires more than identifying individual actions or detecting isolated event boundaries. Complex activities are naturally organized as *partonomies*: fine-grained interactions compose into intermediate events, which in turn form higher-level activities. Recovering this part–whole temporal structure is important for models that must reason about what has happened, how an activity is progressing, and what may occur next. Yet most video understanding methods operate at a single temporal resolution or assume a predefined set of annotation scales, limiting their ability to *discover* hierarchical event structure. Inferring such structure from unlabeled video is challenging, as neither the appropriate representation nor the temporal units are known a priori. Local visual similarity alone is insufficient to determine whether neighboring observations belong to the same higher-level event: visually similar states may participate in different activities, while visually distinct interactions may compose into a coherent temporal unit. Moreover, the relevant notion of similarity changes with abstraction. A boundary that is meaningful at the action level may disappear when reasoning over a broader activity, while distinctions that are weak locally may become important under longer-term temporal context. Consequently, hierarchical event understanding requires not only determining *where* boundaries occur, but also repeatedly redefining *what* constitutes a temporal unit as the level of abstraction increases.

We introduce **RePart**, a framework for unsupervised temporal partonomy induction through *recursive predictive abstraction*. The central idea is that temporal units should not remain fixed across abstraction levels: instead, neighboring units are repeatedly evaluated for compatibility, composed into higher-level events, and then recontextualized by the same predictive mechanism. As illustrated in Fig. 1, RePart uses predictive context to assess how neighboring events evolve and whether they admit a common higher-level explanation. Supported groups are merged into a coarser sequence, after which the predictor is reapplied to the newly induced temporal units. In this way, the representation and the temporal structure co-evolve as the hierarchy is constructed. Unlike predictive hierarchies defined over predefined levels or temporal scales (Mounir et al., 2023b; Chen et al.,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

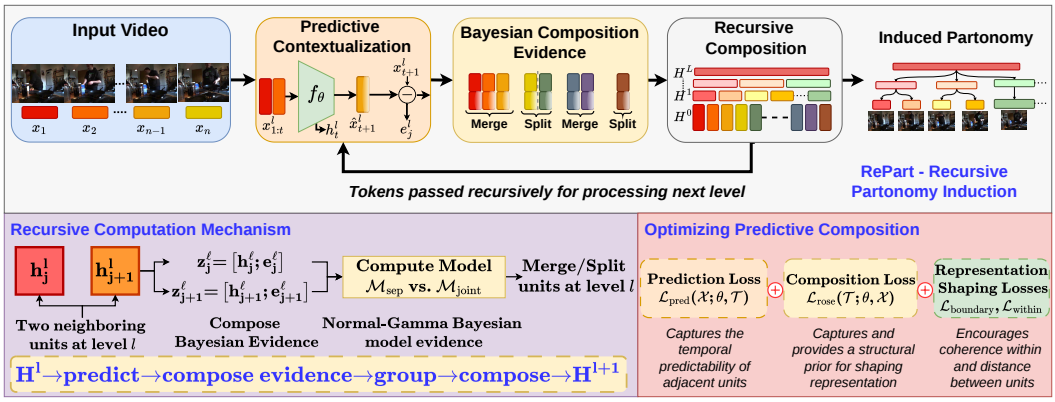


Figure 1: **Overview**. RePart recursively contextualizes temporal units with a shared causal predictor, uses prediction residuals to compute Bayesian evidence for adjacent composition, and merges supported groups into progressively coarser representations. Repeating this process induces a temporal partonomy while predictive and composition-aware losses adapt the underlying representation.

2025b), RePart explicitly changes the units over which subsequent prediction and composition are performed. RePart realizes this idea using a shared causal predictor and Bayesian composition evidence derived jointly from contextualized representations and prediction residuals, recursively producing contiguous higher-level events while preserving explicit part–whole containment. We evaluate RePart on the temporal partonomy benchmark introduced by PARSE (Chen et al., 2025b), spanning Breakfast Actions (Kuehne et al., 2014), 50Salads (Stein & McKenna, 2013), and Assembly101 (Sener et al., 2022). RePart achieves the strongest unsupervised hierarchical F1 on all three datasets and improves over PARSE on both hierarchical metrics across all six dataset–metric comparisons. In particular, hF1 improves from 26.81 to 32.02 on Breakfast-Part, 10.76 to 19.46 on 50Salads-Part, and 21.99 to 25.94 on Assembly101-Part. Ablations show that recursive hierarchy induction benefits from combining contextual state with predictive residuals, and that improvements in hierarchical structure are not explained by fine-scale segmentation accuracy alone. We further evaluate the downstream utility of the induced hierarchy through Assembly101 action anticipation under a controlled TempAgg architecture in which only the hierarchy source is varied. Across the reported horizons, partonomy-conditioned models outperform fixed-scale temporal aggregation, with RePart achieving strong action anticipation and remaining competitive with increasing horizon.

Our **contributions** are three-fold: **(i)** we introduce recursive predictive abstraction, in which temporal units are repeatedly recomposed and recontextualized to induce variable-depth partonomies from unlabeled video; **(ii)** we develop a Bayesian composition criterion, inspired by marginal-likelihood-based hierarchical clustering, that evaluates neighboring units using contextualized state and predictive residuals, together with auxiliary representation-shaping objectives that couple predictive learning with composition-aware structure; and **(iii)** we show that the resulting hierarchies improve structural agreement across three temporal partonomy benchmarks and provide useful hierarchical context for downstream action anticipation.

2 Related Work

Predictive learning and temporal structure. Predictive accounts of perception view internal representations as models of how observations evolve over time, with prediction failures signaling changes in the underlying process (Friston, 2010; Clark, 2013; Rao et al., 2023). Event perception studies similarly associate event boundaries with increases in prediction error and updates to an observer's internal event model (Zacks & Tversky, 2001; Tversky, 2014). Hierarchical predictive inference further suggests that predictive processes operating over different temporal extents can support increasingly abstract representations (Rao et al., 2023; Salvatori et al., 2023; Jiang & Rao, 2023). These ideas have motivated predictive approaches to event segmentation and temporal representation learning in vision (Aakur & Sarkar, 2019; Wang et al., 2024; Aakur & Sarkar, 2020; 2022; Trehan & Aakur, 2024; Mounir et al., 2023b;a), as well as active extensions in which prediction error

guides information acquisition (Trehan & Aakur, 2022; Chen et al., 2025a). Related self-supervised approaches learn temporal representations through predictive or joint-embedding objectives (Assran et al., 2023; Bardes et al., 2023; Littwin et al., 2024; Bardes et al., 2024; Abdelfattah & Alahi, 2024). RePart differs in using prediction not only to contextualize a fixed temporal representation, but also to determine how the temporal units themselves should be recursively composed.

Video event understanding and hierarchical abstraction. Video event understanding has been studied through temporal action segmentation, localization, procedure understanding, and generic event boundary detection. Classical approaches used structured temporal models such as HMMs and CRFs (Kuehne et al., 2020; Chandra et al., 2018), while modern methods employ temporal convolutions (Farha & Gall, 2019; Lea et al., 2017), Transformers (Yi et al., 2021; Fayyaz & Gall, 2020), and graph-based architectures (Huang et al., 2020; Wang et al., 2021). Benchmarks including Breakfast Actions (Kuehne et al., 2014), 50Salads (Stein & McKenna, 2013), EPIC-Kitchens (Damen et al., 2021), Ego4D (Grauman et al., 2022), and Assembly101 (Sener et al., 2022) have supported work on localization (Aakur & Sarkar, 2020; 2022; Trehan & Aakur, 2024; Mounir et al., 2023a), procedure segmentation (Ji et al., 2022; Bansal et al., 2022; 2024; Zhou et al., 2023), and event boundary detection (Shou et al., 2021; Jung et al., 2025; Wang et al., 2022). Hierarchical video representations have also been explored using task annotations, procedural scripts, or predefined temporal scales (Li et al., 2020; Ashutosh et al., 2023; Ghoddoosian et al., 2022; Song et al., 2023; Kung et al., 2025; Flaborea et al., 2024). TW-FINCH (Sarfraz et al., 2021), HVQ (Spurio et al., 2025), and CoSeg (Wang et al., 2024) induce hierarchical or multiscale temporal groupings for unsupervised action segmentation, while Metcalf and Leake (Metcalf & Leake, 2019) learn expectation-driven abstractions across multiple temporal granularities. PARSE (Chen et al., 2025b) more directly studies unsupervised temporal partonomy induction using predictive models, but represents structure over predefined hierarchy levels. RePart instead recursively recomposes adjacent temporal units and reapplies the same predictive mechanism to the newly induced sequence, producing abstraction levels.

Hierarchical composition and Bayesian structure induction. Partonomies describe nested part–whole relations in which lower-level elements compose into progressively larger structures. Part-based representations have a long history in spatial vision (Cheng et al., 2018; Shi et al., 2021; Crivellaro et al., 2018; Heisele et al., 2001; Schneiderman & Kanade, 2004; Huber et al., 2004), while semantic hierarchies and scene graphs capture structured relations among visual entities (Arbeláez et al., 2012; Wang et al., 2018; Krishna et al., 2017; Ji et al., 2020). Our formulation is additionally related to probabilistic hierarchical clustering, where composition is selected using model evidence rather than geometric similarity alone. Bayesian hierarchical clustering evaluates candidate merges through marginal likelihood (Heller & Ghahramani, 2005), while Bayesian Rose Trees generalize this idea to non-binary hierarchical structures using Bayesian evidence for alternative compositions (Blundell et al., 2010). RePart adopts this probabilistic perspective for temporal structure induction, but applies the evidence locally and recursively to contiguous video events whose representations themselves evolve through prediction. The resulting hierarchy is therefore neither a post-hoc clustering of fixed embeddings nor a fixed collection of temporal scales, but a sequence-dependent partonomy jointly shaped by predictive context and recursive composition.

3 REPART: RECURSIVE VIDEO PARTONOMY INDUCTION

Overview. We introduce **RePart**, a framework for inducing temporal partonomies from unlabeled videos through *recursive predictive abstraction*. Given a video $V{=}\{I_i\}_{i=1}^{T}$ represented by features $X{=}\{x_i\}_{i=1}^{T}$, RePart induces a hierarchy $\mathcal{T}{=}\{H^0,\dots,H^L\}$, where each level $H^\ell{=}\{h_j^\ell\}_{j=1}^{N_\ell}$ represents the video at a particular temporal abstraction scale. Each h_j^ℓ spans a contiguous temporal region, with successive levels formed by composing adjacent units so that $N_{\ell+1}{\leq}N_\ell$ and temporal support generally increases with recursion. As illustrated in Fig. 1, a shared predictive model operates on the units at level ℓ and produces next-state residuals that, together with the current-level representations, provide Bayesian evidence for composing neighboring units. The resulting contiguous groups $\mathcal{G}^\ell$ are pooled to form $H^{\ell+1}$, after which the same predictor is reapplied to the newly composed sequence. Prediction therefore informs composition, while composition determines the temporal units over which prediction is performed at the next level. We express this interaction as

$$(\theta_X^\star, \mathcal{T}_X^\star) = \arg\min_{\theta,\mathcal{T}} \mathcal{J}(X;\theta,\mathcal{T}), \tag{1}$$

where $\mathcal{J}$ balances predictive modeling of the current sequence with composition-aware representation learning. In practice, Eq. 1 is not solved by differentiating through $\mathcal{T}$: the hierarchy is instantiated through discrete Bayesian composition decisions under the current representation, while model parameters are optimized using the resulting recursive structure. For each video, RePart first performs sequence-specific unsupervised adaptation and then freezes the adapted model to infer the final hierarchy on the same feature sequence. Unlike looped Transformers (Giannou et al., 2023) that repeatedly refine a fixed tokenization, RePart reapplies shared parameters after recomposing the temporal units, so each recursion operates on a newly induced abstraction level.

3.1 Predictive Contextualization and Recursive Composition

We realize Eq. 1 through two coupled operations: *predictive contextualization*, which models the temporal dynamics at the current abstraction level, and *recursive composition*, which groups adjacent units into higher-level events using their predictive structure. At level ℓ, the shared predictor operates on H^ℓ to produce prediction residuals E^ℓ, which together with the current-level representations define composition evidence and contiguous groups $\mathcal{G}^\ell$. These groups are pooled to form $H^{\ell+1}$, yielding the recursive computation $H^\ell \xrightarrow{\text{predictive contextualization}} E^\ell \xrightarrow{\text{composition evidence}} \mathcal{G}^\ell \xrightarrow{\text{compose}} H^{\ell+1}$. Prediction is therefore conditioned on the units induced at the current level, while composition determines the units presented to the same predictor at the next, allowing RePart to progress from fine temporal structure to increasingly coarse part–whole organization.

Predictive Contextualization. Given $H^\ell = \{h_j^\ell\}_{j=1}^{N_\ell}$, a causal predictor f_θ estimates the next representation as $\hat{h}_{j+1}^\ell = f_\theta(h_{\le j}^\ell)$ for $j = 1, \dots, N_\ell - 1$. We align each prediction error with the unit on which it is observed, $e_1^\ell = \mathbf{0}$ and $e_j^\ell = h_j^\ell - \hat{h}_j^\ell$ for $j = 2, \dots, N_\ell$. The pair (h_j^ℓ, e_j^ℓ) therefore captures both the representation of a temporal unit and its deviation from the preceding predictive context. RePart shares the same predictor across recursive levels (Chen et al., 2025b;a; Aakur & Sarkar, 2022; Mounir et al., 2023b;a); unlike predictive hierarchies defined over fixed temporal scales, however, each application operates on units whose temporal support and granularity have been changed by the preceding composition step. We preserve these predictive dynamics using

$$\mathcal{L}_{\text{pred}}^\ell = \frac{1}{N_\ell - 1} \sum_{j=1}^{N_\ell - 1} \left\| h_{j+1}^\ell - \hat{h}_{j+1}^\ell \right\|_2^2, \tag{2}$$

and aggregate across the induced hierarchy as $\mathcal{L}_{\text{pred}} = \frac{1}{L+1} \sum_{\ell=0}^{L} \mathcal{L}_{\text{pred}}^\ell$. The aligned residuals are then combined with the current-level representations to provide level-specific evidence for composing neighboring temporal units.

Bayesian Composition Evidence. At each level, RePart evaluates whether neighboring temporal units admit a common higher-level explanation or should remain separate. For each position j, we form the joint evidence representation $z_j^\ell = [h_j^\ell, e_j^\ell]$. For adjacent units, we compare a shared explanation against separate explanations using a diagonal Normal–Gamma model,

$$B_j^\ell = \log p\big(z_j^\ell, z_{j+1}^\ell \mid \mathcal{M}_{\text{joint}}\big) - \log p\big(z_j^\ell \mid \mathcal{M}_{\text{sep}}\big) - \log p\big(z_{j+1}^\ell \mid \mathcal{M}_{\text{sep}}\big), \tag{3}$$

where larger values provide greater relative support for composition under the local evidence model. Since evidence magnitudes vary across abstraction levels, we standardize the adjacent scores within each level as $\widetilde{B}_j^\ell = (B_j^\ell - \mu_B^\ell)/(\sigma_B^\ell + \epsilon)$, where μ_B^ℓ and σ_B^ℓ are the empirical mean and standard deviation over adjacent pairs at level ℓ. We then obtain a normalized composition score $p_j^\ell = \sigma(\widetilde{B}_j^\ell)$. Adjacent units are composed when $p_j^\ell \ge \tau$. With the default τ=0.5, this corresponds to non-negative standardized evidence, or equivalently evidence at or above the within-level mean. These decisions define the contiguous composition groups $\mathcal{G}^\ell$ used to construct the next recursive level. By operating on $[H^\ell, E^\ell]$, the criterion reflects both representational compatibility and predictive behavior.

Recursive Composition. The composition scores at level ℓ determine which adjacent temporal units remain within the same higher-level event. RePart places a boundary between units j and $j+1$ whenever $p_j^\ell < \tau$. The remaining adjacent units form contiguous groups $\mathcal{G}^\ell = \{G_k^\ell\}_{k=1}^{N_{\ell+1}}$, where each G_k^ℓ contains one or more consecutive elements of H^ℓ. This preserves temporal order and guarantees that every higher-level unit is composed exclusively from contiguous units at the level below. Each

group is collapsed by mean pooling its current-level representations, $h_k^{\ell+1}=\frac{1}{|G_k^\ell|}\sum_{j\in G_k^\ell} h_j^\ell$. The temporal support of $h_k^{\ell+1}$ is inherited from the union of the spans represented by the elements in G_k^ℓ. The resulting sequence $H^{\ell+1}=\{h_k^{\ell+1}\}_{k=1}^{N_{\ell+1}}$ therefore provides a coarser representation of the same video while preserving explicit parent–child relationships between successive levels. The shared predictor f_θ is then reapplied to $H^{\ell+1}$, producing new residuals and composition evidence for the newly induced abstraction scale. Recursion continues until no further reduction in sequence length is obtained or a predefined maximum depth is reached. The complete output is thus a variable-depth hierarchy in which every higher-level event can be traced to the finer temporal units from which it was recursively composed. During sequence-specific adaptation, the hard composition decisions instantiate the recursive hierarchy used to compute the objective, while gradients update the representation parameters through the predictive and composition-aware losses described below. This adaptation allows the predictive representation and the induced temporal units to co-evolve for the current sequence, refining the features on which subsequent composition decisions are made. The parameters are then frozen, and the same recursive procedure is applied to obtain the final hierarchy.

3.2 Optimizing Predictive Composition

The recursive hierarchy is induced directly from analytic Bayesian composition evidence. During sequence-specific adaptation, the same evidence also provides self-supervision for shaping the predictive representation on which subsequent composition decisions are made. Thus, RePart does not learn a representation first and construct a hierarchy only as post-processing; rather, the hierarchy induced under the current representation determines the recursive units used by the objective, while the objective updates the representation supporting those decisions. At each level ℓ, we optimize

$$\mathcal{L}^\ell = \lambda_P \mathcal{L}^\ell_{\text{pred}} + \lambda_R \mathcal{L}^\ell_{\text{comp}} + \lambda_C \mathcal{L}^\ell_{\text{within}} + \lambda_B \mathcal{L}^\ell_{\text{boundary}}, \tag{4}$$

where the three composition-aware terms respectively distill the analytic Bayesian evidence into an auxiliary compatibility predictor, encourage representational coherence between pairs supported for composition, and preserve separation across pairs supported as boundaries. The objective is averaged over the recursive levels induced for the current video.

Evidence supervision. The standardized Bayesian evidence $\widetilde{B}_j^\ell$ provides an analytically derived target for an auxiliary learned compatibility head. For an adjacent pair $(j, j{+}1)$, the head receives their current-level representations and aligned predictive residuals and outputs a scalar r_j^ℓ. We minimize

$$\mathcal{L}^\ell_{\text{comp}} = \frac{1}{N_\ell - 1} \sum_{j=1}^{N_\ell - 1} \left(r_j^\ell - \widetilde{B}_j^\ell \right)^2. \tag{5}$$

The evidence target is computed from detached representations and residuals, so gradients do not propagate through the Normal–Gamma evidence computation. Note that this is an auxiliary representation-shaping signal; the final recursive hierarchy is induced from the Bayesian evidence.

Within-event coherence. To separate evidence supporting composition from evidence supporting a boundary, we first define $q_j^\ell=\sigma\left(\frac{\widetilde{B}_j^\ell}{\tau_C}\right)$, and construct a positive composition gate $a_j^\ell=\max\big(0,\, 2q_j^\ell - 1\big)$. Thus, only evidence on the composition side of the within-level reference contributes to the attraction term. Let $d_j^\ell=\|h_{j+1}^\ell - h_j^\ell\|_2$. We define

$$\mathcal{L}^\ell_{\text{within}} = \frac{\sum_{j=1}^{N_\ell-1} a_j^\ell (d_j^\ell)^2}{\max\Big(1, \sum_{j=1}^{N_\ell-1} a_j^\ell\Big)}. \tag{6}$$

This encourages neighboring units with stronger relative evidence for composition to remain coherent in the representation space.

Boundary separation. Conversely, we define the complementary boundary gate $w_j^\ell=\max\big(0,\, 1-2q_j^\ell\big)$, so only adjacent pairs lying on the boundary side of the evidence reference contribute to the separation objective. Let $m^\ell=\text{median}_j\, d_j^\ell$ denote the detached median

Method	Sup.	Breakfast-Part		50Salads-Part		Assembly101-Part	
		TED-Sim ↑	hF1 ↑	TED-Sim ↑	hF1 ↑	TED-Sim ↑	hF1 ↑
MS-TCN++ (Li et al., 2023)	Full	40.40	23.36	48.43	32.15	30.67	22.75
FACT (Lu & Elhamifar, 2024)	Full	**48.44**	<u>32.97</u>	**56.14**	<u>44.25</u>	<u>40.96</u>	<u>29.16</u>
DiffACT (Liu et al., 2023)	Full	<u>48.19</u>	**35.66**	<u>55.68</u>	**44.50**	**44.86**	**33.56**
ProTAS (Shen & Elhamifar, 2024)	Full	35.79	19.96	28.24	13.62	34.94	19.45
Hier. Clust. (Chen et al., 2025b)	None	33.35	22.60	**35.59**	<u>17.98</u>	30.20	21.76
STREAMER (Mounir et al., 2023b)	None	18.72	8.68	14.34	3.70	15.04	8.83
PARSE (Chen et al., 2025b)	None	<u>36.04</u>	<u>26.81</u>	20.65	10.76	<u>32.77</u>	<u>21.99</u>
RePart	None	**43.31**	**32.02**	<u>29.33</u>	**19.46**	**36.26**	**25.94**

Table 1: **Partonomy induction performance.** RePart achieves strong unsupervised induction performance across datasets and is competitive with supervised models. The best result for each supervision class is **bolded** and the second best is <u>underlined</u>.

adjacent distance at level ℓ. We optimize

$$\mathcal{L}^{\ell}_{\text{boundary}} = \frac{\sum_{j=1}^{N_\ell - 1} w_j^\ell \left[\max\left(0, m^\ell - d_j^\ell\right)\right]^2}{\max\left(1, \sum_{j=1}^{N_\ell - 1} w_j^\ell\right)}. \tag{7}$$

The level-specific margin adapts the separation scale to the current representation while preventing the model from reducing the loss by changing the margin. The coherence and boundary terms use the Bayesian evidence to attract composition-supported pairs and repel boundary-supported pairs.

Training and Inference. The final sequence-specific objective averages the losses over all induced levels, $\mathcal{J}(X;\theta) = \frac{1}{L+1}\sum_{\ell=0}^{L}\mathcal{L}^\ell$, where the hierarchy $\mathcal{T}_X$ is instantiated by the discrete Bayesian composition procedure under the current representation. Optimization and hierarchy induction are therefore coupled without differentiating through the discrete merge decisions: Bayesian evidence determines the recursive units used by the objective, while the resulting predictive and composition-aware losses update the representation on which subsequent evidence is computed. For each video, RePart performs sequence-specific unsupervised adaptation using only its feature sequence. After adaptation, the parameters are frozen and the recursive procedure is rerun on the same sequence without gradient updates, alternating prediction, Bayesian composition, and pooling from H^0 until no further reduction in sequence length is obtained or the maximum recursion depth is reached. The resulting $\mathcal{T}_X = \{H^0, \ldots, H^L\}$ is therefore a video-specific, variable-depth hierarchy with explicit parent–child containment. While RePart produces variable-depth hierarchies, we export three representative levels from each induced hierarchy, for compatibility with prior work (Mounir et al., 2023b; Chen et al., 2025b). We define a nondegenerate cut as any recursive level containing at least two temporal units. We export the finest and coarsest such cuts and the recursive level nearest their midpoint. We do not use the ground truth for hierarchy induction or export.

Implementation Details. We implement RePart using a causal Transformer with hidden dimension 128, 4 attention heads, 1 Transformer layer, and dropout 0.1. Input video features are ℓ_2 normalized and reduced to 64 dimensions using PCA fit independently to each video. The projected features are mapped to the 128-dimensional predictive space before recursive abstraction. Bayesian composition uses the diagonal Normal–Gamma evidence model described above; when enabled, an additional level-specific PCA is applied only to the detached $[H^\ell, E^\ell]$ evidence representation and does not alter the predictive tokens or pooled representations. Unless otherwise stated, we use composition threshold τ=0.5 and a maximum recursion depth of 8. Each video is initialized independently and adapted for 25 epochs using AdamW with learning rate 3×10^{-4}, weight decay 10^{-4}, and gradient clipping at 1.0. We use loss weights $(\lambda_P, \lambda_R, \lambda_C, \lambda_B) = (1, 1, 1, 0.01)$ and coherence temperature τ_C=2. Code and additional details are provided in the supplementary material.

4 Experimental Evaluation

Datasets. We evaluate RePart on the hierarchical partonomy benchmark introduced by PARSE, constructed from Breakfast Actions (Kuehne et al., 2014), 50Salads (Stein & McKenna, 2013), and Assembly101 (Sener et al., 2022). The benchmark retains videos whose annotations admit a valid

Variant	Boundary F1		Hierarchy	
	Fine ↑	Coarse ↑	TED-Sim ↑	hF1 ↑
$\mathcal{L}_{pred}$ only	28.20	17.28	43.20	31.66
$\mathcal{L}_{pred} + \mathcal{L}_{\text{comp}}$	28.76	15.24	42.66	29.27
$\mathcal{L}_{pred} + \mathcal{L}_{\text{comp}} + \mathcal{L}_{within}$	28.58	20.29	42.10	29.99
$\mathcal{L}_{pred} + \mathcal{L}_{\text{comp}} + \mathcal{L}_{boundary}$	28.06	17.58	43.05	30.21
Full RePart	**29.83**	**22.62**	**43.31**	**32.02**

Table 2: **Objective ablation.** Contribution of the predictive and composition losses to the learned representations, evaluated on Breakfast-Part. Inference still uses Bayesian composition.

Variant	Boundary F1		Hierarchy	
	Fine ↑	Coarse ↑	TED-Sim ↑	hF1 ↑
Fixed-length prediction	30.84	3.65	4.63	4.12
Recursive, H only	27.23	21.37	37.11	28.24
Recursive, E only	29.59	21.24	39.77	30.18
Separate predictor / level	29.63	23.04	42.35	31.84
No predictive adaptation	**30.20**	20.65	43.04	31.76
Full RePart	29.83	**22.62**	**43.31**	**32.02**

Table 3: **Structural ablation.** Effect of recursion, predictive residuals, and predictor sharing, evaluated on Breakfast-Part.

two-level fine-to-coarse temporal hierarchy, yielding complementary cooking and assembly domains with both egocentric and allocentric viewpoints. All methods are evaluated on identical splits and feature representations, following PARSE (Chen et al., 2025b).

Baselines and Metrics. We compare against the supervised temporal segmentation methods MS-TCN++ (Li et al., 2023), FACT (Lu & Elhamifar, 2024), DiffACT (Liu et al., 2023), and ProTAS (Shen & Elhamifar, 2024); non-learned Hierarchical Linkage; the predictive streaming method STREAMER (Mounir et al., 2023b); and PARSE, the strongest prior unsupervised partonomy baseline. Supervised methods are trained independently at each annotation level, while unsupervised baselines construct the hierarchy without ground-truth temporal labels. Following PARSE, we evaluate segmentation quality independently at the fine and coarse levels using mean intersection-over-union (mIoU) and F1, and evaluate hierarchical structure using Tree Edit Distance Similarity (TED-Sim) and hierarchical F1 (hF1). TED-Sim captures agreement in the induced part–whole structure, while hF1 jointly measures structural and temporal agreement across hierarchy levels. Full evaluation details are in the supplementary.

4.1 Partonomy induction

Table 1 presents our main results on hierarchical temporal partonomy induction. RePart achieves the best unsupervised hF1 on all three benchmarks, reaching 32.02 on Breakfast-Part, 19.46 on 50Salads-Part, and 25.94 on Assembly101-Part. Relative to PARSE, the closest prior predictive partonomy model, these correspond to gains of $+5.21$, $+8.70$, and $+3.95$ hF1 points, respectively. RePart also improves TED-Sim over PARSE on every dataset, from 36.04 to 43.31 on Breakfast-Part, 20.65 to 29.33 on 50Salads-Part, and 32.77 to 36.26 on Assembly101-Part. Thus, across the six hierarchy comparisons formed by three datasets and two structural metrics, RePart improves over PARSE in all six cases. The gains are particularly notable because RePart does not rely on temporal annotations or dataset-level hierarchy supervision. On Breakfast-Part, it improves substantially over all prior unsupervised methods on both TED-Sim and hF1, narrowing much of the gap to fully supervised temporal segmentation models. On Assembly101-Part, RePart similarly establishes the strongest unsupervised results on both structural metrics. The behavior on 50Salads-Part is more nuanced: hierarchical clustering retains the highest unsupervised TED-Sim, while RePart attains the highest hF1 and improves markedly over PARSE on both measures. This difference is consistent with the complementary nature of the metrics: TED-Sim emphasizes agreement in hierarchical structure, whereas hF1 additionally rewards temporal alignment of the induced part–whole relationships. More broadly, the results support RePart's central design. Prior predictive approaches such as STREAMER and PARSE model temporal structure at fixed or predefined abstraction scales, whereas RePart repeatedly composes the units themselves and reapplies the same predictive mechanism to the resulting coarser sequence. The consistent improvements in TED-Sim and hF1 indicate that this recursive change in representational unit is useful for recovering coherent part–whole organization, rather than merely improving isolated boundary decisions at individual scales. In particular, the largest improvement over PARSE occurs on 50Salads-Part, where RePart gains $+8.68$ TED-Sim and $+8.70$ hF1, suggesting that recursive composition can recover substantially stronger hierarchy even when flat temporal segmentation remains challenging. RePart is also competitive with fully supervised methods despite operating without temporal labels. On Breakfast-Part, its hF1 of 32.02 approaches FACT (32.97) and is substantially higher than MS-TCN++ and ProTAS; on Assembly101-Part, RePart exceeds MS-TCN++ and ProTAS in hF1.

Method	Sup.	Breakfast-Part				50Salads-Part				Assembly101-Part			
		Fine		Coarse		Fine		Coarse		Fine		Coarse	
		mIoU ↑	F1 ↑	mIoU ↑	F1 ↑	mIoU ↑	F1 ↑	mIoU ↑	F1 ↑	mIoU ↑	F1 ↑	mIoU ↑	F1 ↑
MS-TCN++ (Li et al., 2023)	Full	9.84	18.84	6.04	11.38	5.73	10.88	4.36	8.54	4.86	9.43	2.94	5.80
FACT (Lu & Elhamifar, 2024)	Full	**12.92**	**23.40**	7.29	12.18	**8.61**	**15.90**	7.81	14.34	**15.72**	**27.42**	4.33	8.45
DiffACT (Liu et al., 2023)	Full	11.86	22.27	**8.49**	**14.43**	8.05	14.87	**7.95**	**14.63**	13.02	23.18	**5.03**	**9.64**
ProTAS (Shen & Elhamifar, 2024)	Full	11.75	22.15	6.32	11.74	5.51	10.51	3.65	7.08	11.84	21.50	3.31	6.50
Hier. Clust. (Chen et al., 2025b)	None	16.36	28.77	8.62	16.35	4.00	7.67	3.97	7.63	**20.42**	**34.20**	**5.78**	**11.15**
STREAMER (Mounir et al., 2023b)	None	11.13	21.05	4.18	9.74	4.35	8.39	1.44	2.84	11.71	21.10	1.88	3.69
PARSE (Chen et al., 2025b)	None	**18.93**	**32.49**	10.98	19.48	**5.10**	**10.21**	4.40	8.57	18.34	31.10	5.34	10.44
RePart	None	15.87	29.83	**12.12**	**22.62**	3.78	8.40	**4.88**	**10.10**	19.63	33.08	5.40	10.27

Table 4: **Fine- and coarse-grained segmentation performance.** RePart leads in unsupervised coarse segmentation for Breakfast-Part and 50Salads-Part, while remaining competitive across fine-grained and Assembly101-Part segmentation. Supervised methods are included as reference points.

4.2 What Drives Recursive Partonomy Induction?

Objective components. Table 2 examines how the predictive and composition objectives contribute to RePart. Prediction alone already provides a strong structural signal, reaching 43.20 TED-Sim and 31.66 hF1, suggesting that temporal predictability is itself informative for discovering event organization. However, its coarse boundary F1 is substantially lower than the full model (17.28 vs. 22.62). The individual composition terms affect different aspects of the representation: adding within-event coherence produces the largest improvement in coarse F1 among the partial objectives (20.29), while boundary separation better preserves hierarchical agreement. Neither term alone, nor composition supervision by itself, consistently improves all metrics. Their combination in full RePart yields the strongest overall result, improving both fine and coarse boundary F1 while attaining the best TED-Sim and hF1. This suggests that the composition losses act jointly: Bayesian evidence provides the structural signal, while within-event coherence and boundary separation shape the representation on opposite sides of a prospective composition decision.

Recursive structure and predictive evidence. Table 3 shows that recursion is the dominant structural ingredient. Fixed-length prediction retains strong fine boundary F1 (30.84) but fails to recover coarse structure (3.65 F1) or a meaningful hierarchy (4.12 hF1). Recursive composition restores hierarchy quality, with predictive residuals alone outperforming state alone (30.18 vs. 28.24 hF1), and their combination giving the strongest TED-Sim and hF1 (43.31 and 32.02). Using a separate predictor at each level performs similarly but slightly worse structurally, supporting parameter sharing as a compact mechanism for reapplying prediction as the temporal units change. Notably, removing sequence-specific predictive adaptation leaves TED-Sim nearly unchanged (43.04 vs. 43.31) but reduces coarse F1 (20.65 vs. 22.62), suggesting that recursive Bayesian composition provides much of the structural inductive bias, while adaptation primarily helps coarse event organization.

Hierarchy	α=1 ↑	α=2 ↑	α=3 ↑	α=4 ↑	α=5 ↑
TempAgg	8.60	6.60	5.73	4.62	4.26
MS-TCN++	9.12	7.47	6.51	5.83	5.35
FACT	9.02	7.39	6.01	5.52	5.31
DiffACT	8.28	7.10	6.46	5.91	4.93
Hier. Clust.	9.65	7.62	6.42	5.88	5.05
PARSE	9.41	**7.80**	6.80	6.20	5.42
RePart	**10.15**	7.54	**6.98**	**6.36**	**5.71**

Table 5: **Action anticipation on Assembly101.** Action T5R across increasing anticipation horizons (denoted by α). All methods use the same TempAgg architecture; only the temporal hierarchy is varied.

Downstream Utility for Action Anticipation. We evaluate whether the induced partonomy is useful beyond segmentation, using Assembly101 action anticipation with TempAgg (Sener et al., 2022) as a controlled baseline. For each anticipation horizon, the partonomy is constructed using only observations available up to the corresponding cutoff frame; no method has access to future frames. We fix the features, architecture, training, and anticipation protocol, and vary only the supplied temporal hierarchy. This provides a direct test of whether the organization of the observed prefix, rather than additional visual information or model capacity, contributes useful predictive context. Table 5 shows that RePart's recursively induced hierarchy supports prediction across multiple timescales. RePart achieves the best Action T5R at four of five horizons (α=1, 3, 4, 5) and is second to PARSE at α=2. Its advantage persists at longer horizons, reaching 5.71 at α=5 compared with 5.42 for

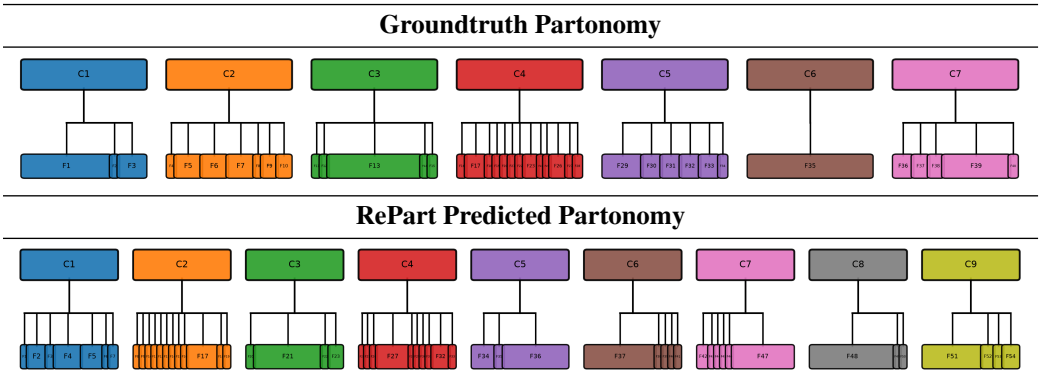


Figure 2: **Qualitative temporal partonomy induction** on Breakfast Actions (making a sandwich). RePart recovers a coarse-to-fine hierarchy that broadly reflects the ground-truth part–whole structure. C1, C2, . . . and F1, F2, . . . denote sequential coarse- and fine-scale temporal units.

PARSE and 4.26 for fixed-scale TempAgg. This is particularly relevant at larger anticipation horizons, where immediate local evidence is less informative, and prediction must rely more heavily on the broader temporal organization of the observed activity. The results suggest that the induced hierarchy functions as a compact summary of the observations that supports action anticipation.

Fine- and Coarse-Grained Segmentation. Table 4 shows that RePart's gains are concentrated at coarser scales. It achieves the strongest unsupervised coarse segmentation on Breakfast-Part and 50Salads-Part, improving PARSE from 10.98 to 12.12 mIoU and 19.48 to 22.62 F1 on Breakfast-Part, and from 4.40 to 4.88 mIoU and 8.57 to 10.10 F1 on 50Salads-Part. Fine-scale results are mixed: RePart improves over PARSE on Assembly101-Part but trails it on Breakfast-Part and 50Salads-Part. Thus, the hierarchy gains in Table 1 are not explained by uniformly better segmentation at every scale. Together with Sec. 4.2, these results indicate that recursive Bayesian composition provides the primary structural inductive bias, while predictive adaptation and composition-aware representation shaping mainly refine higher-level event organization. RePart improves part–whole consistency by repeatedly redefining the temporal units, rather than by optimizing boundaries alone.

Qualitative Discussion. Figure 2 shows a representative Breakfast Actions example in which RePart recovers much of the ground-truth coarse-to-fine organization for making a sandwich. The main failure mode arises from the irreversibility of hard recursive composition: incorrect merges at finer levels cannot be undone at later stages, while fragmented fine-level structure tends to persist upward, producing over-segmentation across multiple abstraction levels. This also highlights a limitation of evaluating a variable-depth hierarchy at fixed benchmark scales, since local grouping errors can propagate through the recursion and affect the exported fine/coarse levels.

5 Conclusion and Future Work

We presented **RePart**, an unsupervised framework for temporal partonomy induction through recursive predictive abstraction. RePart repeatedly recomposes temporal units using Bayesian evidence over current representations and prediction residuals, then reapplies a shared causal predictor to the newly induced sequence, yielding variable-depth hierarchies with explicit part–whole containment. Across Breakfast Actions, 50Salads, and Assembly101, RePart improves hierarchical agreement over prior unsupervised methods while remaining competitive on fine- and coarse-grained segmentation, and controlled anticipation experiments show that the induced hierarchies provide useful downstream temporal context. Ablations indicate that recursive Bayesian composition provides the primary structural inductive bias, while predictive adaptation and composition-aware objectives refine higher-level organization. The main limitation is the irreversibility of hard composition: early grouping errors can propagate upward and produce persistent over-segmentation. Future work will explore recoverable composition, richer variable-depth evaluation, and intervention-based analysis of whether modifying intermediate event structure causally affects downstream predictions.

REFERENCES

Sathyanarayanan Aakur and Sudeep Sarkar. Action localization through continual predictive learning. In *ECCV*, pp. 300–317. Springer, 2020.

Sathyanarayanan Aakur and Sudeep Sarkar. Actor-centered representations for action localization in streaming videos. In *ECCV*, pp. 70–87. Springer, 2022.

Sathyanarayanan N Aakur and Sudeep Sarkar. A perceptual prediction framework for self supervised event segmentation. In *CVPR*, pp. 1197–1206, 2019.

Mohamed Abdelfattah and Alexandre Alahi. S-JEPA: A joint embedding predictive architecture for skeletal action recognition. In *ECCV*, pp. 367–384. Springer, 2024.

Pablo Arbeláez, Bharath Hariharan, Chunhui Gu, Saurabh Gupta, Lubomir Bourdev, and Jitendra Malik. Semantic segmentation using regions and parts. In *2012 CVPR*, pp. 3378–3385. IEEE, 2012.

Kumar Ashutosh, Rohit Girdhar, Lorenzo Torresani, and Kristen Grauman. HierVL: Learning hierarchical video-language embeddings. In *CVPR*, pp. 23066–23078, 2023.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *CVPR*, pp. 15619–15629, 2023.

Siddhant Bansal, Chetan Arora, and CV Jawahar. My view is the best view: Procedure learning from egocentric videos. In *ECCV*, pp. 657–675. Springer, 2022.

Siddhant Bansal, Chetan Arora, and CV Jawahar. United we stand, divided we fall: Unitygraph for unsupervised procedure learning from videos. In *WACV*, pp. 6509–6519, 2024.

Adrien Bardes, Jean Ponce, and Yann LeCun. MC-JEPA: A joint-embedding predictive architecture for self-supervised learning of motion and content features. *arXiv preprint arXiv:2307.12698*, 2023.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. V-JEPA: Latent video prediction for visual representation learning, 2024.

Charles Blundell, Yee Whye Teh, and Katherine A. Heller. Bayesian rose trees. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, UAI'10, pp. 65–72, Arlington, Virginia, USA, 2010. AUAI Press. ISBN 9780974903965.

Siddhartha Chandra, Camille Couprie, and Iasonas Kokkinos. Deep spatio-temporal random fields for efficient video segmentation. In *CVPR*, pp. 8915–8924, 2018.

Zhou Chen, Sanjoy Kundu, Harsimran S Baweja, and Sathyanarayanan N Aakur. Ease: Embodied active event perception via self-supervised energy minimization. *IEEE Robotics and Automation Letters*, 2025a.

Zhou Chen, Joe Lin, and Sathyanarayanan N. Aakur. Generalized event partonomy inference with structured hierarchical predictive learning. *arXiv preprint arXiv:2512.04219*, 2025b.

Jingchun Cheng, Yi-Hsuan Tsai, Wei-Chih Hung, Shengjin Wang, and Ming-Hsuan Yang. Fast and accurate online video object segmentation via tracking parts. In *CVPR*, pp. 7415–7424, 2018.

Andy Clark. Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3):181–204, 2013.

Alberto Crivellaro, Mahdi Rad, Yannick Verdie, Kwang Moo Yi, Pascal Fua, and Vincent Lepetit. Robust 3D object tracking from monocular images using stable parts. *TPAMI*, 40(6):1465–1479, 2018.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The EPIC-KITCHENS dataset: Collection, challenges and baselines. *TPAMI*, 43(11):4125–4141, 2021.

Yazan Abu Farha and Jurgen Gall. MS-TCN: Multi-stage temporal convolutional network for action segmentation. In *CVPR*, pp. 3575–3584, 2019.

Mohsen Fayyaz and Jurgen Gall. SCT: Set constrained temporal transformer for set supervised action segmentation. In *CVPR*, pp. 501–510, 2020.

Alessandro Flaborea, Guido Maria D'Amely Di Melendugno, Leonardo Plini, Luca Scofano, Edoardo De Matteis, Antonino Furnari, Giovanni Maria Farinella, and Fabio Galasso. Prego: online mistake detection in procedural egocentric videos. In *CVPR*, pp. 18483–18492, 2024.

Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11 (2):127–138, 2010.

Reza Ghoddoosian, Saif Sayed, and Vassilis Athitsos. Hierarchical modeling for task recognition and action segmentation in weakly-labeled instructional videos. In *WACV*, pp. 1922–1932, 2022.

Angeliki Giannou, Shashank Rajput, Jy-yong Sohn, Kangwook Lee, Jason D Lee, and Dimitris Papailiopoulos. Looped transformers as programmable computers. In *International Conference on Machine Learning*, pp. 11398–11442. PMLR, 2023.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4D: Around the world in 3,000 hours of egocentric video. In *CVPR*, pp. 18995–19012, 2022.

Bernd Heisele, Thomas Serre, Massimiliano Pontil, Thomas Vetter, and Tomaso A Poggio. Categorization by learning and combining object parts. In *NeurIPS*, pp. 1239–1245, 2001.

Katherine A Heller and Zoubin Ghahramani. Bayesian hierarchical clustering. In *Proceedings of the 22nd international conference on Machine learning*, pp. 297–304, 2005.

Yifei Huang, Yusuke Sugano, and Yoichi Sato. Improving action segmentation via graph-based temporal reasoning. In *CVPR*, pp. 14024–14034, 2020.

Daniel Huber, Anuj Kapuria, Raghavendra Donamukkala, and Martial Hebert. Parts-based 3D object classification. In *CVPR*, volume 2, pp. II–II. IEEE, 2004.

Jingwei Ji, Ranjay Krishna, Li Fei-Fei, and Juan Carlos Niebles. Action genome: Actions as compositions of spatio-temporal scene graphs. In *CVPR*, pp. 10236–10247, 2020.

Lei Ji, Chenfei Wu, Daisy Zhou, Kun Yan, Edward Cui, Xilin Chen, and Nan Duan. Learning temporal video procedure segmentation from an automatically collected large dataset. In *WACV*, pp. 1506–1515, 2022.

Linxing Preston Jiang and Rajesh PN Rao. Dynamic predictive coding explains both prediction and postdiction in visual motion perception. In *Annual Meeting of the Cognitive Science Society*, volume 45, 2023.

Hyungrok Jung, Daneul Kim, Seunggyun Lim, Jeany Son, and Jonghyun Choi. Online generic event boundary detection. In *ICCV*, pp. 13741–13750, 2025.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *IJCV*, 123(1):32–73, 2017.

Hilde Kuehne, Ali Arslan, and Thomas Serre. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *CVPR*, pp. 780–787, 2014.

Hilde Kuehne, Alexander Richard, and Juergen Gall. A hybrid RNN-HMM approach for weakly supervised temporal action segmentation. *TPAMI*, 42(4):765–779, 2020.

Chi-Hsi Kung, Frangil Ramirez, Juhyung Ha, Yi-Ting Chen, David Crandall, and Yi-Hsuan Tsai. What changed and what could have changed? state-change counterfactuals for procedure-aware video representation learning. In *ICCV*, pp. 12294–12306, 2025.

Colin Lea, Michael D Flynn, Rene Vidal, Austin Reiter, and Gregory D Hager. Temporal convolutional networks for action segmentation and detection. In *CVPR*, pp. 156–165, 2017.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. HERO: Hierarchical encoder for video+language omni-representation pre-training. In *EMNLP*, pp. 2046–2065, 2020.

Shijie Li, Yazan Abu Farha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. MS-TCN++: Multi-stage temporal convolutional network for action segmentation. *TPAMI*, 45(6):6647–6658, 2023. doi: 10.1109/TPAMI.2020.3021756.

Etai Littwin, Omid Saremi, Madhu Advani, Vimal Thilak, Preetum Nakkiran, Chen Huang, and Joshua Susskind. How JEPA avoids noisy features: The implicit bias of deep linear self distillation networks. *NeurIPS*, 37:91300–91336, 2024.

Daochang Liu, Qiyue Li, Anh-Dung Dinh, Tingting Jiang, Mubarak Shah, and Chang Xu. Diffusion action segmentation. In *ICCV*, pp. 10139–10149, 2023.

Zijia Lu and Ehsan Elhamifar. Fact: Frame-action cross-attention temporal modeling for efficient action segmentation. In *CVPR*, pp. 18175–18185, 2024.

Katherine Metcalf and David Leake. Unsupervised hierarchical temporal abstraction by simultaneously learning expectations and representations. In *IJCAI*, pp. 3144–3150, 2019.

Ramy Mounir, Ahmed Shahabaz, Roman Gula, Jörn Theuerkauf, and Sudeep Sarkar. Towards automated ethogramming: Cognitively-inspired event segmentation for streaming wildlife video monitoring. *IJCV*, 131(9):2267–2297, 2023a. doi: 10.1007/s11263-023-01781-2.

Ramy Mounir, Sujal Vijayaraghavan, and Sudeep Sarkar. Streamer: Streaming representation learning and event segmentation in a hierarchical manner. *NeurIPS*, 36:45694–45715, 2023b.

Rajesh PN Rao, Dimitrios C Gklezakos, and Vishwas Sathish. Active predictive coding: A unifying neural model for active perception, compositional learning, and hierarchical planning. *Neural Computation*, 36(1):1–32, 2023.

Tommaso Salvatori, Ankur Mali, Christopher L Buckley, Thomas Lukasiewicz, Rajesh PN Rao, Karl Friston, and Alexander Ororbia. Brain-inspired computational intelligence via predictive coding. *arXiv preprint arXiv:2308.07870*, 2023.

Saquib Sarfraz, Naila Murray, Vivek Sharma, Ali Diba, Luc Van Gool, and Rainer Stiefelhagen. Temporally-weighted hierarchical clustering for unsupervised action segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11225–11234, 2021.

Henry Schneiderman and Takeo Kanade. Object detection using the statistics of parts. *IJCV*, 56(3): 151–177, 2004.

Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *CVPR*, pp. 21096–21106, 2022.

Yuhan Shen and Ehsan Elhamifar. Progress-aware online action segmentation for egocentric procedural task videos. In *CVPR*, pp. 18186–18197, 2024.

Shaoshuai Shi, Zhe Wang, Jianping Shi, Xiaogang Wang, and Hongsheng Li. From points to parts: 3D object detection from point cloud with part-aware and part-aggregation network. *TPAMI*, 43 (8):2647–2664, 2021. doi: 10.1109/TPAMI.2020.2977026.

Mike Zheng Shou, Stan Weixian Lei, Weiyao Wang, Deepti Ghadiyaram, and Matt Feiszli. Generic event boundary detection: A benchmark for event segmentation. In *ICCV*, pp. 8075–8084, 2021.

Yale Song, Eugene Byrne, Tushar Nagarajan, Huiyu Wang, Miguel Martin, and Lorenzo Torresani. Ego4D goal-step: Toward hierarchical understanding of procedural activities. *NeurIPS*, 36: 38863–38886, 2023.

Federico Spurio, Emad Bahrami, Gianpiero Francesca, and Juergen Gall. Hierarchical vector quantization for unsupervised action segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 6996–7005, 2025.

Sebastian Stein and Stephen J McKenna. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 729–738, 2013.

Shubham Trehan and Sathyanarayanan N Aakur. Towards active vision for action localization with reactive control and predictive learning. In *WACV*, pp. 783–792, 2022.

Shubham Trehan and Sathyanarayanan N Aakur. Self-supervised multi-actor social activity understanding in streaming videos. In *ICPR*, pp. 293–309. Springer, 2024.

Barbara Tversky. Where partonomies and taxonomies meet. In *Meanings and Prototypes (RLE Linguistics B: Grammar)*, pp. 334–344. Routledge, 2014.

Panqu Wang, Pengfei Chen, Ye Yuan, Ding Liu, Zehua Huang, Xiaodi Hou, and Garrison Cottrell. Understanding convolution for semantic segmentation. In *WACV*, pp. 1451–1460. IEEE, 2018.

Tao Wang, Ning Xu, Kean Chen, and Weiyao Lin. End-to-end video instance segmentation via spatial-temporal graph neural networks. In *ICCV*, pp. 10797–10806, 2021.

Xiao Wang, Jingen Liu, Tao Mei, and Jiebo Luo. CoSeg: Cognitively inspired unsupervised generic event segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 35(9):12507–12517, 2024.

Yuxuan Wang, Difei Gao, Licheng Yu, Weixian Lei, Matt Feiszli, and Mike Zheng Shou. Geb+: A benchmark for generic event boundary captioning, grounding and retrieval. In *ECCV*, pp. 709–725. Springer, 2022.

Fangqiu Yi, Hongyu Wen, and Tingting Jiang. ASFormer: Transformer for action segmentation. In *BMVC*, 2021. doi: 10.5244/C.35.49.

Jeffrey M Zacks and Barbara Tversky. Event structure in perception and conception. *Psychological Bulletin*, 127(1):3, 2001.

Honglu Zhou, Roberto Martín-Martín, Mubbasir Kapadia, Silvio Savarese, and Juan Carlos Niebles. Procedure-aware pretraining for instructional video understanding. In *CVPR*, pp. 10727–10738, 2023.

LLM Use Acknowledgement. We used ChatGPT 5.6 to polish the writing, verify the mathematical formulation, and generate the visualization code for qualitative results.

A Exporting Variable-Depth Hierarchies to Fixed Levels

RePart induces a variable-depth hierarchy whose number of recursive levels can differ across videos, whereas the PARSE evaluation protocol expects three predicted hierarchy levels. We therefore use an annotation-free export procedure that maps each induced hierarchy to three representative cuts. First, we order all recursive cuts from fine to coarse and discard terminal degenerate cuts containing fewer than two segments. From the remaining cuts, we select the finest nondegenerate cut, the coarsest nondegenerate cut, and an intermediate cut corresponding to the recursive level nearest the midpoint between them. If three or fewer eligible cuts are available, all available cuts are retained. This "spread" strategy samples the full fine-to-coarse extent of the predicted hierarchy rather than simply taking the first three recursion levels. The selected cuts are exported to the legacy L2/L3/L4 interface only for compatibility with the existing evaluation code; these names do not imply correspondence to particular semantic scales. Boundaries are obtained directly from the start of each predicted segment after the first and converted to the temporal indexing expected by the benchmark. No ground-truth boundaries, segment counts, action labels, or evaluation scores are used during hierarchy induction or level selection. During evaluation only, the three anonymous exported cuts are matched to the benchmark's annotated fine and coarse levels subject to their temporal ordering, following prior works (Mounir et al., 2023b; Chen et al., 2025b). We consider the valid ordered correspondences between exported scales and report the benchmark metrics under the corresponding fine–coarse assignment, following the existing evaluation protocol. Thus, ground truth is used only to evaluate and align the exported anonymous scales, never to construct the hierarchy or determine which recursive levels are exported.

Export-level correspondence. As a sanity check on the annotation-free export procedure, we record which exported recursive levels are selected during benchmark evaluation. On Breakfast-Part, fine matches L3 for $42/48$ videos (87.5%) and L2 for $6/48$ (12.5%), while coarse matches L4 for $42/48$ (87.5%) and L3 for $6/48$ (12.5%). On 50Salads-Part, fine matches L3 for $14/20$ videos (70%) and L2 for $6/20$ (30%), whereas coarse matches L4 for $17/20$ videos (85%) and L3 for $3/20$ (15%). On Assembly101-Part, fine matches L3 for $194/200$ videos (97%) and L2 for $6/200$ (3%), while coarse matches L4 for $194/200$ videos (97%) and L3 for $6/200$ (3%). Overall, the benchmark correspondence is strongly concentrated on adjacent exported levels, indicating that the annotation-free export preserves a stable fine-to-coarse ordering while allowing limited flexibility in effective hierarchy depth.

A.1 Evaluation Metrics

We evaluate both the quality of temporal segmentation at individual hierarchy levels and the consistency of the induced partonomy across levels. For each annotated scale, we report mean intersection-over-union (mIoU) and F1. mIoU measures the temporal overlap between predicted and ground-truth segments, while F1 summarizes the precision–recall tradeoff of predicted event boundaries within the benchmark tolerance. Higher values indicate better agreement with the annotated segmentation. To evaluate the hierarchy itself, we additionally report Tree Edit Distance Similarity (TED-Sim) and hierarchical F1 (hF1). TED-Sim measures the structural similarity between the predicted and reference partonomies; higher values indicate fewer edits are required to transform one hierarchy into the other. hF1 instead measures agreement in the hierarchical relationships while accounting for temporal alignment, with higher values indicating better preservation of the annotated part–whole structure. These metrics are complementary: fine- and coarse-level mIoU/F1 assess segmentation quality at fixed scales, whereas TED-Sim and hF1 measure whether those segmentations form a coherent hierarchy. A method can therefore achieve similar per-level segmentation accuracy while differing substantially in how consistently the levels are organized into a partonomy.

B Additional Implementation Details

Input preprocessing. RePart operates on pre-extracted video features and does not update the visual backbone. For each video independently, features are ℓ_2 normalized and reduced to 64 dimensions

using PCA fit only to that video's unlabeled feature sequence. The resulting features are projected to the model hidden dimension before recursive processing. Neither the input PCA nor any learned model parameters are transferred between videos.

Predictive model. The causal predictor is a 1-layer Transformer with hidden dimension 128, 4 attention heads, and dropout 0.1. The same predictor parameters are reused at every recursive level. Causal attention is implemented using scaled dot-product attention with `is_causal=True`. The feed-forward block has dimension 4×128 with GELU activation. At each level, the predictor operates on the currently induced temporal units rather than the original fixed tokenization.

Bayesian composition. Composition evidence is computed from the detached concatenation of each current-level representation and its aligned prediction residual. We use a diagonal Normal–Gamma model with $\kappa_0 = 0.1$ and $\alpha_0 = 2$, setting the prior mean to the empirical mean of the current level and the scale parameter from its empirical variance. Adjacent log-evidence scores are standardized independently within each recursive level and clipped to $[-8, 8]$ before conversion to composition scores. Unless otherwise noted, adjacent units are merged at threshold $\tau = 0.5$, corresponding to evidence at or above the within-level mean. Recursion terminates when the sequence is no longer reduced, fewer than two units remain, or a maximum depth of 8 is reached.

Sequence-specific optimization. Each video is processed independently. We initialize a new RePart model, optimize it for 25 epochs on that video's unlabeled feature sequence using AdamW with learning rate 3×10^{-4}, weight decay 10^{-4}, and gradient clipping at 1.0. One optimization step constructs the complete recursive hierarchy under the current parameters, computes the objective over all instantiated levels, averages the level losses, and performs a single parameter update. Hard Bayesian composition decisions are not differentiated through. After optimization, the model is frozen and the hierarchy is recomputed from the same sequence without gradient updates. The model and per-video PCA are then discarded.

Objective and inference. We use $(\lambda_P, \lambda_R, \lambda_C, \lambda_B) = (1, 1, 1, 0.01)$ and coherence temperature $\tau_C = 2$ for the full model. The analytic Bayesian evidence, rather than the auxiliary compatibility head, determines all composition decisions at inference. Reported hierarchies are exported using the annotation-free procedure described in Sec. A.

C Hyperparameter sensitivity

Figure 3 evaluates the sensitivity of RePart to the four objective weights and coherence temperature on Breakfast-Part. For each hyperparameter, we marginalize over the remaining sweep configurations and report the mean and standard deviation of fine/coarse F1, TED-Sim, and hF1. Overall, RePart is stable across the tested ranges: the marginal mean varies by at most 0.40 points in TED-Sim and 0.48 points in hF1 across any individual hyperparameter. Fine-scale F1 is particularly insensitive, with a maximum variation of only 0.12 points. The larger, although still modest, variation occurs in coarse F1, indicating that the representation-shaping terms primarily affect higher-level temporal organization rather than fine boundary localization.

Prediction weight λ_P. Changing λ_P from 1 to 2 has essentially no effect on any metric: fine and coarse F1 vary by only 0.03 and 0.05 points, respectively, while TED-Sim and hF1 change by 0.15 and 0.26 points. This suggests that once next-state prediction is present in the objective, the induced hierarchy is not sensitive to its precise relative weighting.

Composition weight λ_R. The auxiliary evidence-supervision weight is similarly stable over $\lambda_R \in \{0.5, 1\}$. TED-Sim and hF1 vary by only 0.12 and 0.25 points, while the largest change is a 0.41-point difference in coarse F1. This is consistent with the role of $\mathcal{L}_{\text{comp}}$ as a representation-shaping signal rather than the mechanism directly responsible for composition: hierarchy induction itself continues to use the analytic Bayesian evidence.

Boundary weight λ_B. Performance is nearly invariant over the tested $\lambda_B \in \{0.001, 0.01, 0.05\}$ range. The marginal means vary by only 0.08 points in fine F1, 0.21 in coarse F1, 0.07 in TED-Sim, and 0.11 in hF1. Thus, boundary separation provides a useful regularizing signal without requiring precise tuning of its magnitude.

Coherence weight λ_C. Among the loss weights, λ_C has the largest effect, although the overall variation remains small. Across $\lambda_C \in \{0, 0.25, 0.5, 1\}$, TED-Sim changes by 0.23 points and hF1

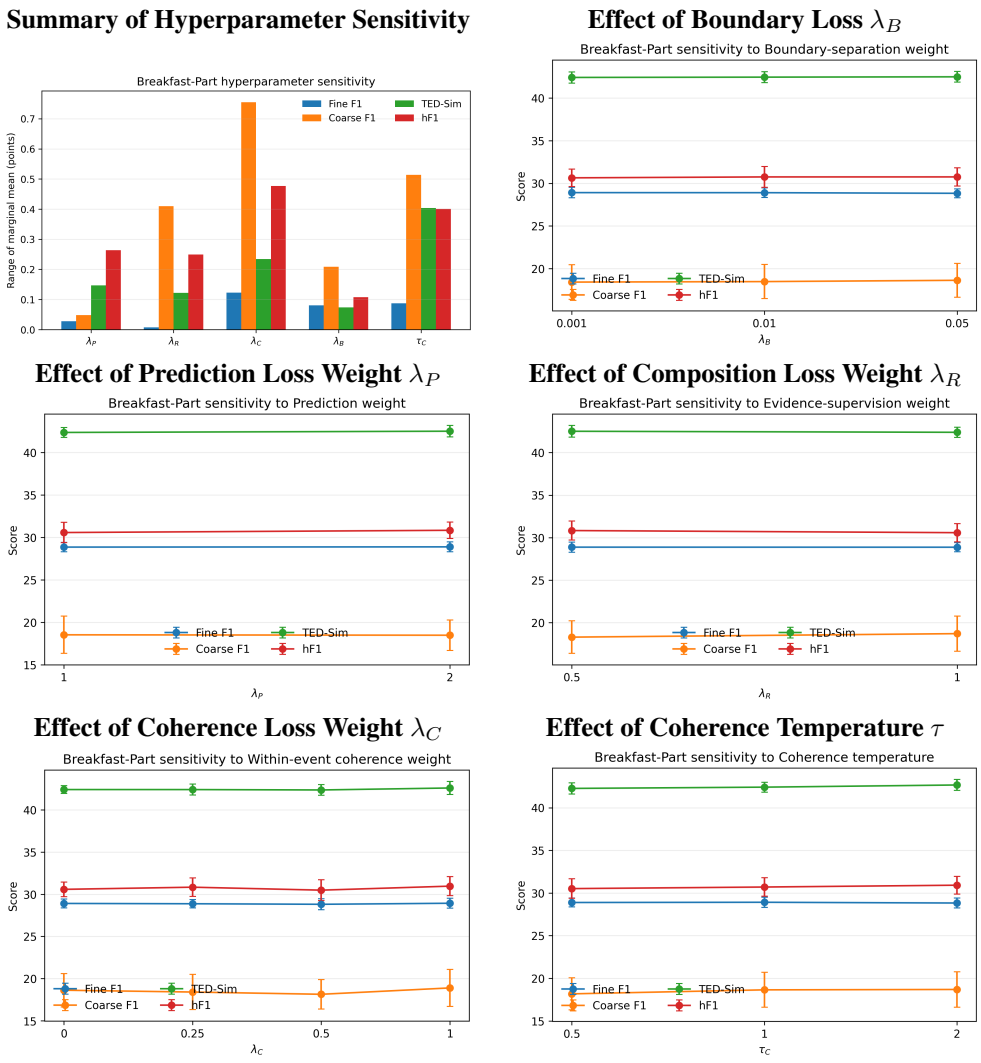


Figure 3: **Ablation studies** to evaluate the sensitivity of RePart to different hyperparameter values on Breakfast-Part.

by 0.48, while coarse F1 varies by 0.75 points. The primary effect is therefore at the coarser temporal scale, consistent with within-event coherence encouraging recursively composed units to form more stable higher-level events. Importantly, no narrow optimum is observed: hierarchy quality remains similar across the full tested range.

Coherence temperature τ_C. Varying $\tau_C \in \{0.5, 1, 2\}$ likewise produces only modest changes, with ranges of 0.09, 0.51, 0.40, and 0.40 points for fine F1, coarse F1, TED-Sim, and hF1, respectively. This indicates that RePart is relatively insensitive to the sharpness with which standardized Bayesian evidence is converted into the coherence and boundary gates.

Taken together, these sweeps indicate that RePart's performance is governed primarily by the recursive composition mechanism rather than a narrowly tuned combination of objective coefficients. The representation-shaping hyperparameters mainly modulate coarse-scale organization, while the induced hierarchy remains stable over broad settings.

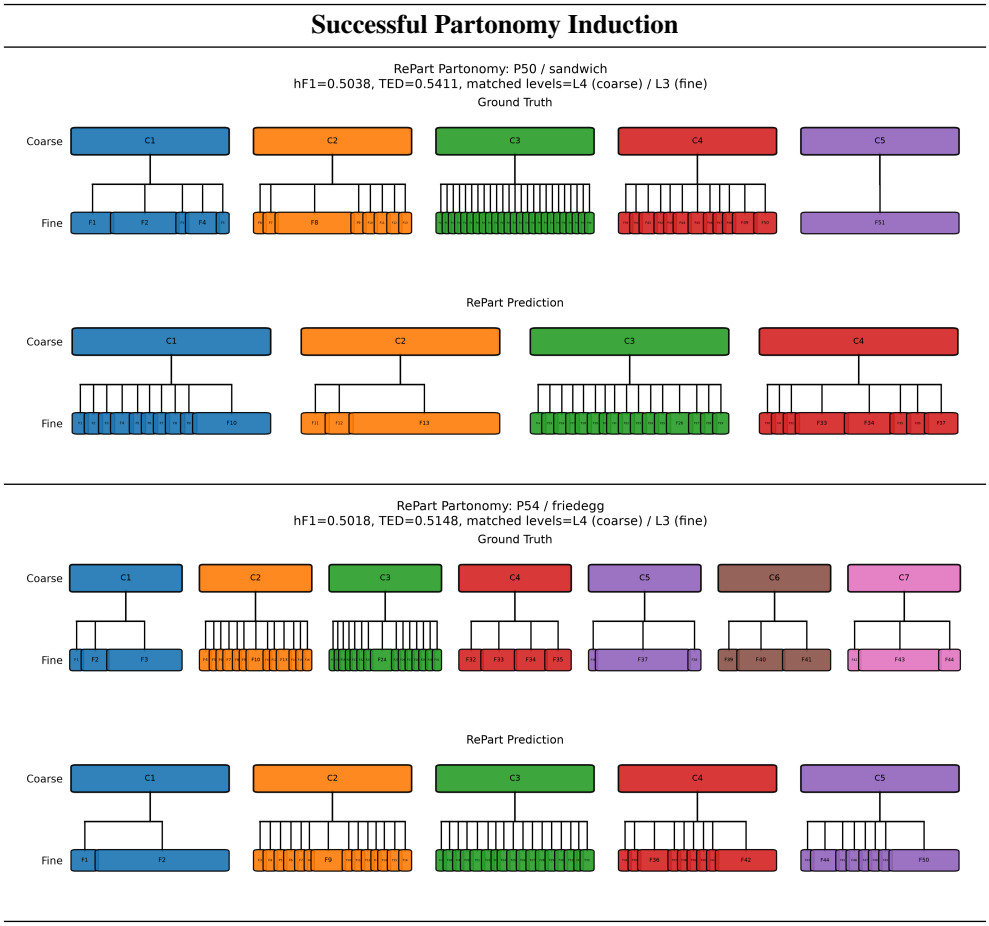


Figure 4: **Qualitative Visualization of successful partonomy induction** by RePart on two representative videos from Breakfast Actions.

D Additional Qualitative Visualization

Figures 4 and 5 illustrate representative successful and unsuccessful Breakfast-Part hierarchies. In the successful cases, RePart recovers the dominant coarse event ordering and assigns most fine-scale units to the correct parent events, even when the predicted number or exact placement of fine segments differs from the annotation. This supports the quantitative result that hierarchy quality can remain strong without requiring perfect frame-level segmentation. The failure cases reveal a complementary limitation. RePart tends to over-segment the activity, producing too many coarse units and fragmented fine structure relative to the reference hierarchy. Because composition is recursive and hard, early fragmentation is inherited by subsequent levels and cannot be merged back into an alternative finer interpretation once the hierarchy has advanced. The examples therefore reinforce two observations from the quantitative analysis: RePart is effective when local predictive evidence supports stable part–whole grouping, but errors in early grouping can propagate upward and reduce structural agreement even when local temporal boundaries remain plausible.

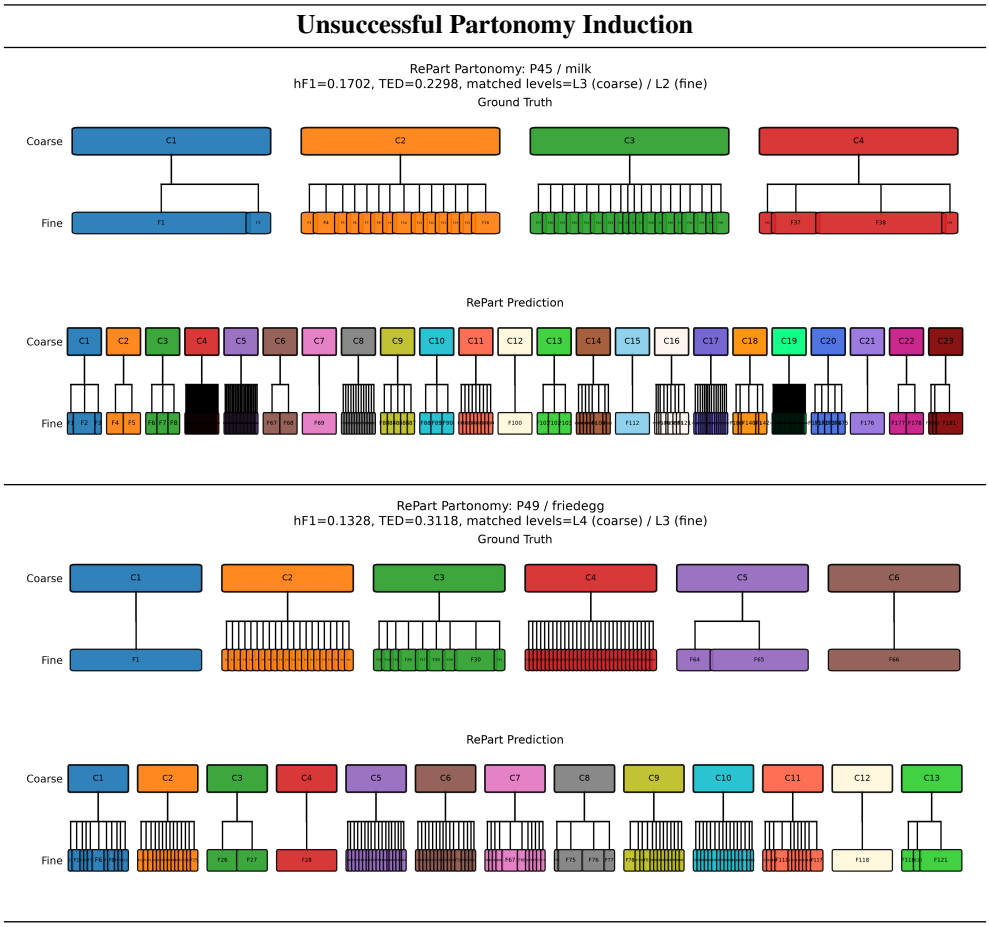


Figure 5: **Qualitative Visualization of unsuccessful partonomy induction** by RePart on two representative videos from Breakfast Actions.